Zero-shot Image-to-Image Translation

Abstract

Large-scale text-to-image generative models have shown their remarkable ability to synthesize diverse, high-quality images. However, directly applying these models for real image editing remains challenging for two reasons. First, it is hard for users to craft a perfect text prompt depicting every visual detail in the input image. Second, while existing models can introduce desirable changes in certain regions, they often dramatically alter the input content and introduce unexpected changes in unwanted regions. In this work, we introduce pix2pix-zero, an image-to-image translation method that can preserve the original image's content without manual prompting. We first automatically discover editing directions that reflect desired edits in the text embedding space. To preserve the content structure, we propose cross-attention guidance, which aims to retain the cross-attention maps of the input image throughout the diffusion process. Finally, to enable interactive editing, we distill the diffusion model into a fast conditional GAN. We conduct extensive experiments and show that our method outperforms existing and concurrent works for both real and synthetic image editing. In addition, our method does not need additional training for these edits and can directly use the existing pre-trained text-to-image diffusion model.